ARMY RESEARCH LABORATORY

Arabic Optical Character Recognition (OCR) Evaluation in Order to Develop a Post-OCR Module

by Brian Kjersten

ARL-MR-0798 September 2011

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-MR-0798 **September 2011**

Arabic Optical Character Recognition (OCR) Evaluation in Order to Develop A Post-OCR Module

Brian Kjersten
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	***Form Approved* *OMB No. 0704-0188***

1. REPORT DATE *(DD-MM-YYYY)* September 2011	**2. REPORT TYPE** Final	**3. DATES COVERED (From - To)**

4. TITLE AND SUBTITLE Arabic Optical Character Recognition (OCR) Evaluation in Order to Develop a Post-OCR Module	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Brian Kjersten	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory ATTN: RDRL-CII-T 2800 Powder Mill Road Adelphi, MD 20783-1197	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-MR-0798
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Optical character recognition (OCR) is the process of converting an image of a document into text. While progress in OCR research has enabled low error rates for English text in low-noise images, performance is still poor for noisy images and documents in other languages. We intend to create a post-OCR processing module for noisy Arabic documents which can correct OCR errors before passing the resulting Arabic text to a translation system. To this end, we are evaluating an Arabic-script OCR engine on documents with the same content but varying levels of image quality. We have found that OCR text accuracy can be improved with different stages of pre-OCR image processing: (1) filtering out low-contrast images to avoid "hallucination" of characters, (2) removing marks from images with cleanup software to prevent their misrecognition, and (3) zoning multi-column images with segmentation software to enable recognition of all zones. The specific errors observed in OCR will form the basis of training data for our post-OCR correction module.

15. SUBJECT TERMS

OCR, Arabic, noise

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Brian Kjersten
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	18	**19b. TELEPHONE NUMBER** (*Include area code*) (301) 394-1165

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction/Background

Optical character recognition (OCR) is the process of identifying the text in an image and saving the text characters in an electronic file. The U.S. Army Research Laboratory (ARL) has incorporated foreign-language OCR as a front end in its machine translation (MT) system, the Forward Area Language Converter (FALCon) since 1996 (*3, 4*). The FALCon, shown in figure 1, enables the user to scan documents into the system and get a simple translation. Because of the limitations of OCR and MT technology, the translations are not expected to be perfect. However, the translations are good enough to enable basic keyword identification, so that the user may decide whether to send the document to a linguist to translate later (*9*). With the goal of improving such systems, we focus in this report on evaluating OCR for a large-scale, noisy document collection, as a first step in developing a post-OCR correction module for FALCon-like translation workflows.

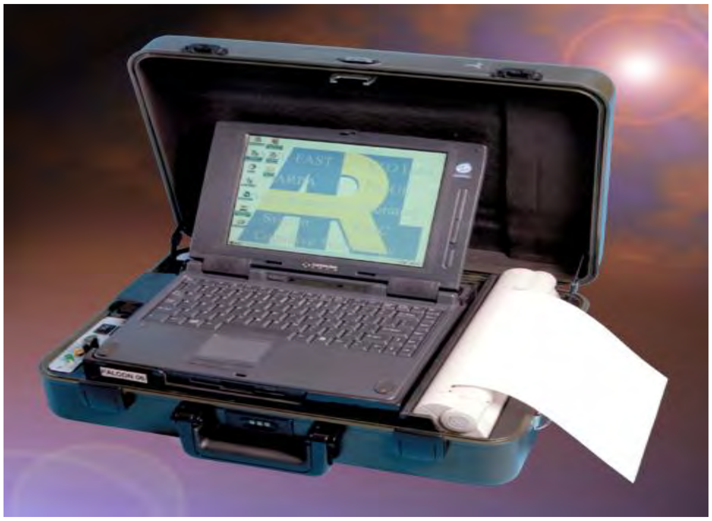

Figure 1. A photograph of the portable FALCon system, with a laptop, a scanner, and a battery pack.

ARL has access to an electronic collection of approximately two million images of Arabic documents. This collection is particularly challenging because it is very noisy and diverse. A few examples of documents in our collection are shown in figure 2.

Besides being low-resolution, the images have large amounts of added image noise. The collection contains several different types of documents, including tables, forms, reports, and letters. Many of these are completely handwritten, and many more have some handwriting in the margins. Some images are blurred or faded to the point of illegibility. Others are mostly or completely black.

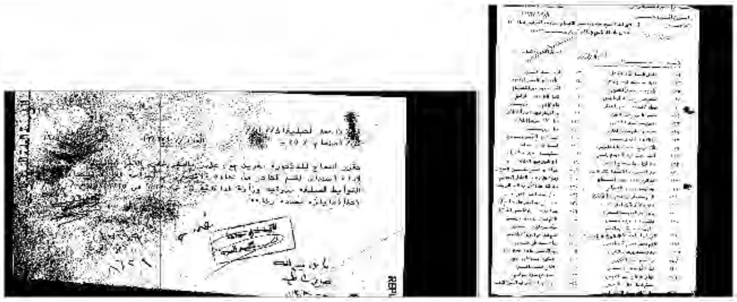

Figure 2. Examples of some of the noisy Arabic documents in our collection.

With the current state of the technology, English language OCR can obtain high character accuracy rates in low-noise situations. However, performance is much worse when trying to recognize a foreign language, especially a language which does not use the Latin alphabet (*8*). The Arabic alphabet is especially difficulty because some Arabic characters are more easily confusable than Latin characters, and because character boundaries are not as obvious as they are in English (*2,* 8).

The presence of image noise is also detrimental to OCR. Removal of noise from images is an active field of research. Current research identifies several types of noise, including clutter, which refers to black spots with width greater than the text width; lines, which have the same width as text, but are much longer than the text width; and speckle, which consists of spots on the order of the same size as the text width. Speckle is especially damaging to Arabic text, because many letters are distinguished only by the presence and number of dots. In addition, a document can have blur, pixel shift, or bleed-through, which are nonlinearly dependent on the content of the document (*1*).

We intend to create a post-OCR processing module for noisy Arabic documents which can correct OCR errors before passing the resulting Arabic text to a translation system. To this end, we are evaluating OCR engines on documents with the same content and varying levels of image quality. The specific errors observed in OCR will form the basis of training data for a post-OCR correction module.

2. Experiment

A diagram of our experiment is shown in figure 3. The first step was to measure the accuracy of our OCR system on our collection, and to determine the effect of image cleaning software on OCR performance. We constructed three versions of a 117 image subset of the full collection for our tests: the original documents, the cleaned documents, and synthesized documents. The original documents are the 117 images without any modification. To produce the cleaned documents, we ran an image cleanup program called ArtClean which removes some of the image noise. The synthesized documents were produced from the human-generated typed transcriptions of the original documents, saved as images. By scoring the performance of the original documents, we learn how accurately our OCR system performs on noisy images. Scores for the cleaned images help us identify which types of noise interfere the most with recognition. The synthesized images give us an upper bound on the performance of our current OCR system because they do not have image noise.

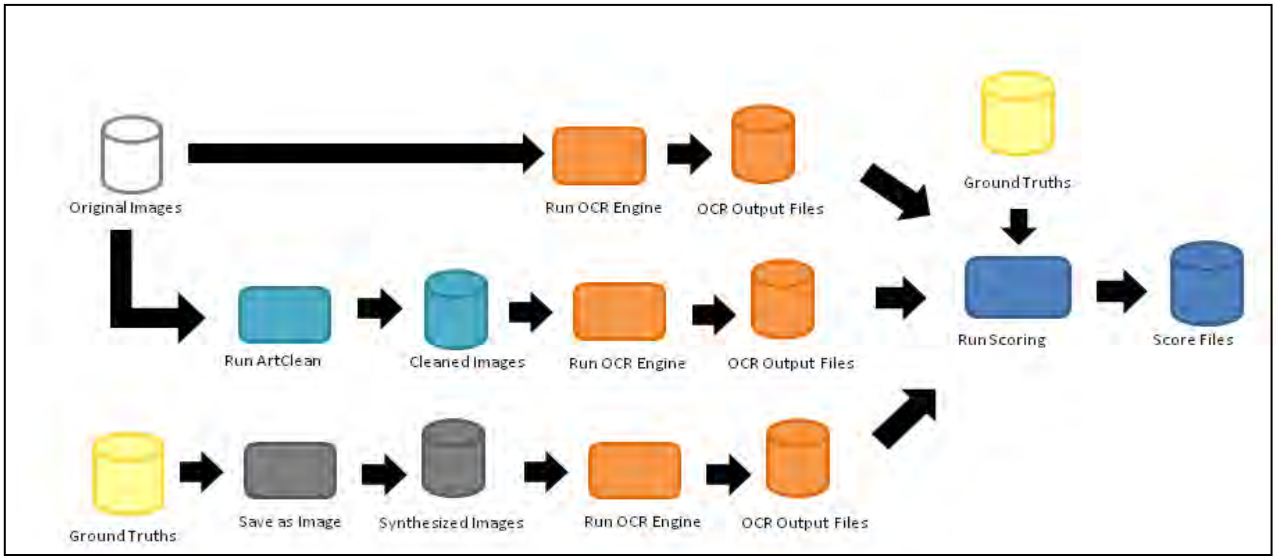

Figure 3. The workflow of our experiment. We conduct OCR on three versions of the same documents: (i) the original noisy images, (ii) the images after cleaning, and (iii) images synthesized by saving ground truth text files as images. We then score the OCR outputs with a character-by-character comparison to the ground truth files to compute OCR accuracy.

For evaluating the accuracy of OCR output, we used a scoring program which compares a manually-produced ground truth file character-by-character with the OCR output. This program determines the most likely alignment between text strings, and counts the number of insertion errors, deletion errors, and substitution errors in the OCR results. It then calculates an accuracy score using these error counts and the total number of characters in the ground truth file using the following equation:

$$a = (NC - NE) / NC \qquad (1)$$

In the above, a is the accuracy score, NC is the number of characters in the ground truth file, and NE is the number of errors. Using this program, we were able to establish a baseline of how well our Arabic OCR currently works.

The scoring program also produces detailed reports specifying the exact errors that occur in the text. Using this, we will derive the statistics necessary for a simple channel model for our post-OCR module.

3. Results and Discussion

Table 1 presents the character error rates we found on the three versions of our dataset.

Table 1. The character accuracies of our documents according to our scoring software. Subset A and Subset B are different folders created by the original owners of the collection.

	Complete set **117 docs.**	**Subset A** **37 docs.**	**Subset B** **80 docs.**
Original	22.37%	23.65%	21.45%
Cleaned	21.80%	12.07%	28.79%
Synthesized	60.02%	70.10%	60.29%

We were initially quite surprised to see that cleaning the images degraded the overall average character accuracy rate of the full set. After inspecting the files individually, however, we found that cleaning improved the performance on some documents, and worsened the performance on others. Notably, for nine of the original documents, the OCR was unable to recognize any characters and returned a blank text file which yielded an accuracy score of 0.00%. When OCR is conducted on those same documents after being cleaned, the OCR "hallucinated" many characters, sometimes many more characters than the ground truth document contained, yielding a negative accuracy score.

When we manually remove these nine documents from the full set, the average accuracy of the original documents improves from 22.37% to 24.50%. and the average accuracy of the cleaned documents improves from 21.80% to 31.83%. Visually inspecting the images, these images look like they are more "shaded" than any other image in the 117-image subset because of a large amount of salt and pepper noise across the page. This suggests that performance can be improved significantly by using a different cleaning algorithm to handle these low-contrast images. Overall precision can also be improved by filtering out low-contrast documents of this sort at the cost of reduced recall. The distribution of the character accuracy is seen in figure 4.

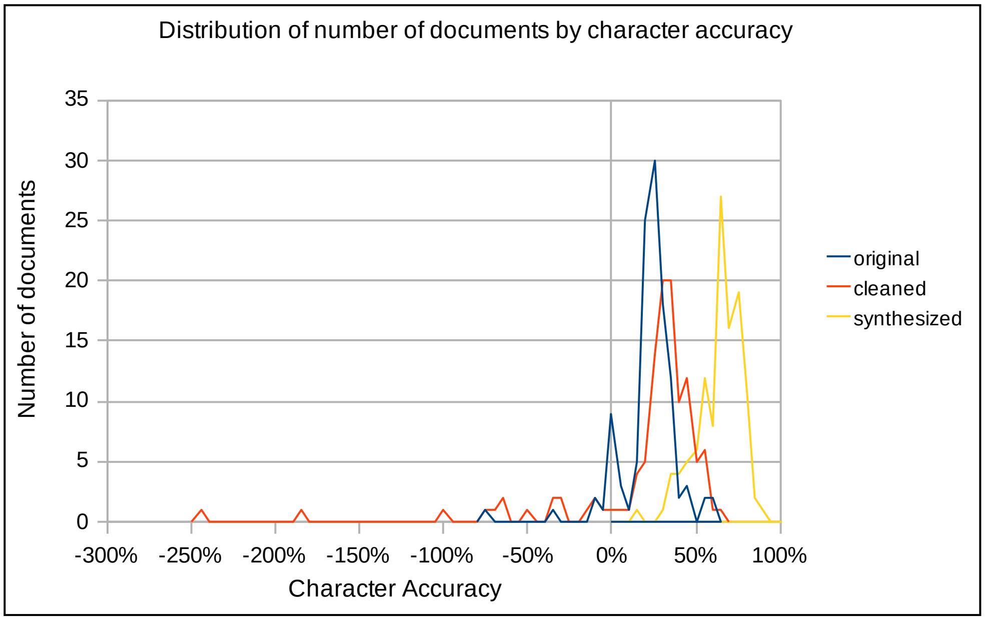

Figure 4. Histogram of character accuracies for the documents.

The cleaned documents have a peak accuracy than is higher than that of the original documents, and the synthesized documents have a peak accuracy much higher than both. This suggests that the cleaned documents tend to perform better than the original documents, but several outliers lower average score of the cleaned documents. Figure 4 highlights the fact that there are nine original documents that have a accuracy scores of 0.00%, and more cleaned documents that have accuracy scores below 0%. We looked at the documents which had high scores and low scores, to determine which characteristics (such as layout, format, or noise) were significantly different at each end of the scoring scale, and to identify the types of noise that caused the most problems.

Other than the nine documents which had scores of exactly 0%, the original documents with low scores all had poor image quality, with various black marks interfering with the text. The high scoring ones all had comparatively good image quality, with very little speckle relative to the other documents. This shows that image quality is a very significant factor in determining OCR performance in this set. These classes of documents are shown in figure 5.

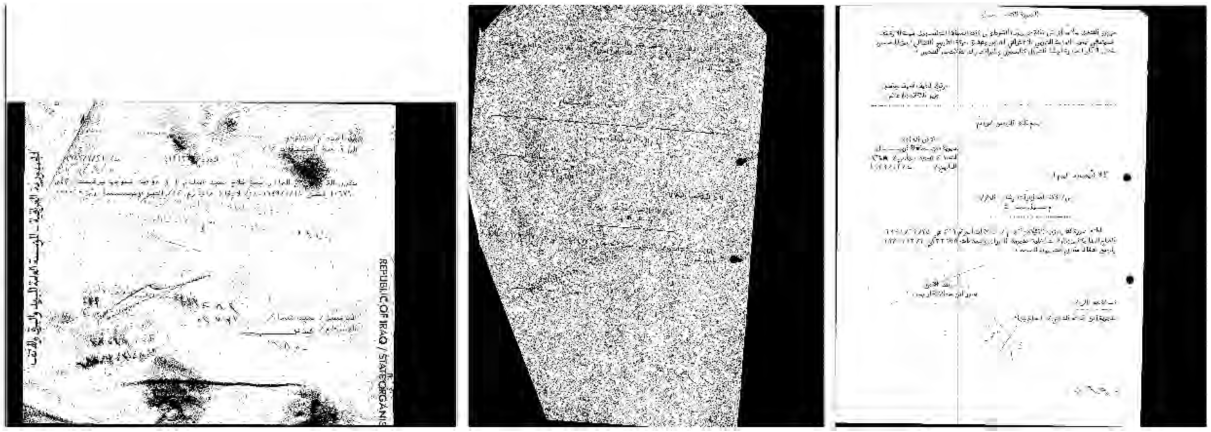
REPUBLIC OF IRAQ / STATE ORGANIS

Figure 5. The image on the left is an original document which received a very low score, the image in the middle received a score of zero, and the image on the right received a comparatively high score.

We wanted to get a better sense of how the OCR results of individual documents were affected by the cleaning process. Figure 6 shows the accuracy scores of each document in its original, cleaned, and synthesized forms, with the negative scores removed from the chart to give a closer view of the positive range.

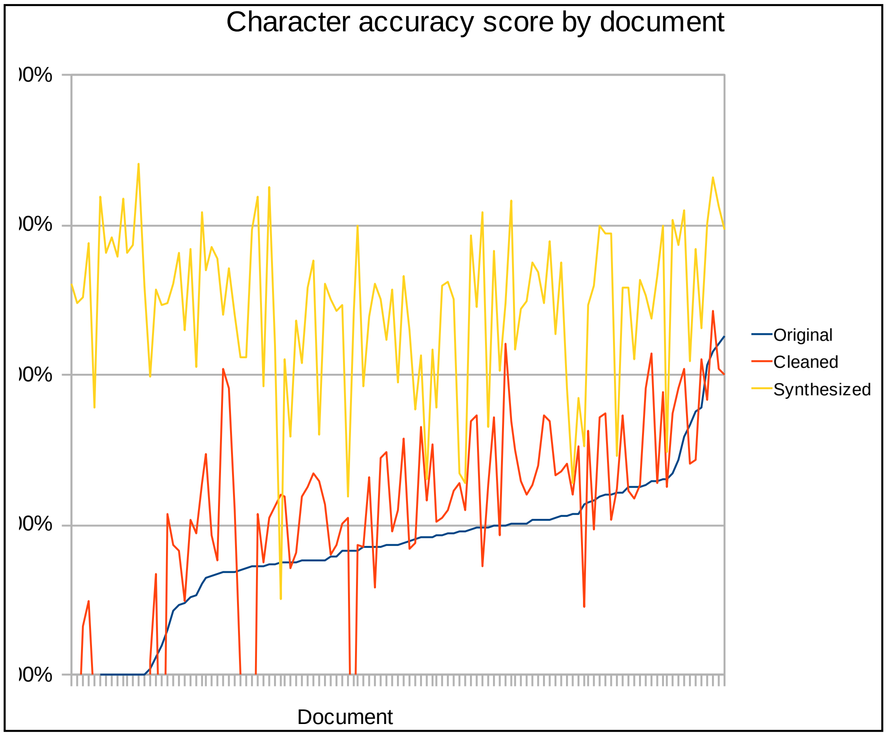

Figure 6. The character accuracies of each document in its original, cleaned, and synthesized form. The documents are sorted by the score of the original. For simplicity, accuracies below 0% are not shown.

We found that 84 of the 117 documents were improved by cleaning, 32 were worsened, and one had no change in accuracy score. This shows that our current cleaning software is effective in improving document quality. With the observations already noted, we are able to identify additional areas of improvement that can be pursued.

Although noise was clearly a very significant factor in determining the scores, we wanted to explore what other factors may exist. We examined the performance of the synthesized set to identify other limitations in the software. The worst-scoring of the synthesized documents had two columns. It was clear from the OCR output that this layout was not supported by the OCR, because the second column was absent entirely from the OCR output. In the future, we will test for improvements in OCR performance by adding a document segmentation module as a preprocessing step before the OCR. The segmenter will separate the document into zones for independent OCR processing. At this point, we do not know the accuracy of the segmenter or how it will impact the OCR accuracy.

In addition, character error rates for the synthesized images seem to be high enough to make translation difficult. This suggests that we add a post-OCR module to clean up the text before sending it to a machine translation system.

The proposed workflow is shown in figure 7.

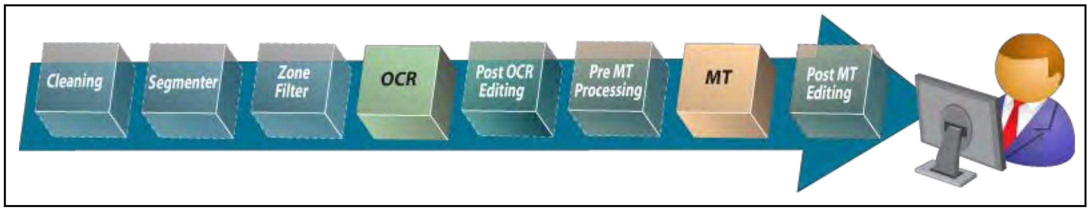

Figure 7. Proposed workflow, with all modules needed for end-to-end image processing, OCR, and MT.

In our experiments, we have used the cleaning module to remove noise from the image prior to running the OCR module. We have shown that the cleaning module is useful, and have identified areas of improvement for it. It is also clear that we will need to assess adding a segmenter module into the workflow for handling multi-column documents. These results support our plan to add a post-OCR editing module into the workflow to repair the text, because even with synthesized images, the character error rate is high enough to cause difficulty for a machine translation system.

4. Conclusion and Future Work

We have established a baseline measurement of how well our existing optical character recognition technology works on a subset of our full collection, both with and without cleaning. Our collection of Arabic documents has a very high rate of optical character recognition errors.

We have found that image quality is the largest factor in determining OCR performance. Our current pre-OCR image cleanup software is good in that it improves OCR performance on most documents. However, our image cleaning software worsens character accuracy on low-contrast images. Therefore, it would be reasonable to filter out low-contrast images from the collection or develop our cleaning system to better handle low-contrast images. Our system also performs poorly on images with a large amount of marks on them, so that would be a worthy focus area for pre-OCR processing work. In addition to image quality, the document format is a significant factor. Our OCR software does not support multi-column documents. In order to handle these documents, we intend to explore adding a document segmentation module to our workflow.

We plan to explore the possibility of "repairing" the Arabic text as output by the OCR with a post-OCR processing module that will use the noisy channel model. The noisy channel model is a statistical representation of language or any other signal. It is widely used in speech recognition, and has been applied for OCR error correction (*5, 7*). In the noisy channel model, we treat the characters as though they are generated by a (language) source, in such a way that the probability of each character depends on the other characters that are present. The character signal is then randomly corrupted by a channel. Here, the statistical source is a mathematical approximation of the writer, and the channel is an approximation of everything that occurs in scanning a document and trying to recognize characters. It is our task to develop a decoder which can reconstruct the original true characters based on the characters which are observed at the output of the OCR system (*5*). For our purposes, the true signal, denoted T, is the actual sequence of characters in the document. The observed signal, denoted O, is the output of an OCR system which has errors. T' is our decoder's best guess of what characters were in the original document. The most likely true data is

$$\begin{aligned} T' &= \text{argmax } Pr(T|O) \\ &= \text{argmax } Pr(O|T)\, Pr(T) / Pr(O) \\ &= \text{argmax } Pr(O|T)\, Pr(T) \end{aligned}$$

The function Pr(O|T) is called the channel model, and it tells the probability of every kind of error. The channel model probabilities can be estimated automatically using a set of OCRed documents for which we have a reliable transcription. Training for the channel model consists of counting how many times each true character is realized as each particular observed character, and comparing that to the number of times that true character occurs. In addition, it is necessary to do additional calculations called smoothing to account for the fact that even events with zero counts in the training data have nonzero probability. It is often necessary to take into account sets of two or more characters because characters can be recognized differently based on the context (*6*).

The function P(T) is called the source model or the language model. It is a function which assigns a probability to every string of characters that can be generated by the source, and it is effectively our way of approximating the actions of a writer using a statistical model. Unlike the channel model, the language model can be trained on any set of Arabic data, because it is independent of the OCR problem. Luckily, we have access to large collections of data in Arabic, including the Arabic Gigaword Corpus and transcriptions of some of our Arabic documents. Language models can calculate probabilities on the character level or on the word level. Very simple language models do not consider the context of a symbol at all. Those language models are called unigram language models because they consider single symbols in isolation. Slightly more complex language models operate on the assumption that the probability of a word depends on the words that precede it. These may calculate the probability of a word based on the one, two, or more words which precede it. These language models are called bigram, trigram, and N-gram language models respectively.

In general, language models can take into account much more nuanced information than only a few preceding symbols. Significantly, they can take into account more complex linguistic information such as what part of speech is expected in that location, whether the word is more likely to be a subject or an object, whether the word needs to agree in person, number, or gender with another word, whether a word is likely to be a pronoun that is coreferenced with another word, and the location of a word within a phrase. Such linguistic information is often more relevant to Arabic than it is to English, because Arabic has more features such as agreement. We say that Arabic is more "morphologically rich" than English. We intend to develop OCR correction modules using an N-gram language model and a linguistically informed language model to compare their effects on character accuracy of the images.

5. References

1. Agrawal, M.; Doermann, D. *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, 2009, 556–560.
2. Bazzi, I.; Schwartz, R.; Makhoul, J. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1999**, *21*, 495–504.
3. Fisher, F.; Voss, C. *Proceedings of the Advanced Information Processing and Analysis Conference*, 1997.
4. Fisher, F.; Schlesiger, C.; Decrozant, L.; Zuba, R.; Holland, M.; Voss, C. R. *Proceedings of the Advanced Information Processing and Analysis Conference*, 1999.
5. Jelinek, F. *Statistical Methods for Speech Recognition*, The MIT Press, 1997.
6. Kolak, O.; Resnick, R.; Byrne, W. *Proceedings of the 2003 Symposium on Document Image Understanding Technology*, 2003, 313–317.
7. Kolak, O.; Resnick P. *Proceedings of the 2nd International Conference on Human Language Technology Research*, 2002, 257–262.
8. Lorigo, L. M.; Govindaraju, V. *IEEE Transactions on Pattern Analysis Machine Intelligence* **2006**, *28*, 712–24.
9. Voss, C.; Van Ess-Dykema, C. ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems, 2000, 1–8.

NO. OF
COPIES ORGANIZATION

1 ADMNSTR
ELECT DEFNS TECHL INFO CTR
ATTN DTIC OCP
8725 JOHN J KINGMAN RD STE 0944
FT BELVOIR VA 22060-6218

1 CD OFC OF THE SECY OF DEFNS
ATTN ODDRE (R&AT)
THE PENTAGON
WASHINGTON DC 20301-3080

1 US ARMY RSRCH DEV AND ENGRG CMND
ARMAMENT RSRCH DEV & ENGRG CTR
ARMAMENT ENGRG & TECHNLGY CTR
ATTN AMSRD AAR AEF T
J MATTS
BLDG 305
ABERDEEN PROVING GROUND MD 21005-5001

1 US ARMY INFO SYS ENGRG CMND
ATTN AMSEL IE TD A RIVERA
FT HUACHUCA AZ 85613-5300

1 COMMANDER
US ARMY RDECOM
ATTN AMSRD AMR
W C MCCORKLE
5400 FOWLER RD
REDSTONE ARSENAL AL
35898-5000

NO. OF
COPIES ORGANIZATION

1 US GOVERNMENT PRINT OFF
DEPOSITORY RECEIVING SECTION
ATTN MAIL STOP IDAD J TATE
732 NORTH CAPITOL ST NW
WASHINGTON DC 20402

1 DIRECTOR
US ARMY RSRCH LAB
ATTN RDRL ROE V W D BACH
PO BOX 12211
RESEARCH TRIANGLE PARK NC 27709

11 US ARMY RSRCH LAB
ATTN IMNE ALC HRR MAIL & RECORDS MGMT
ATTN RDRL CII T
M HOLLAND
C VOSS (5 COPIES)
B KJERSTEN
ATTN RDRL CIO LL TECHL LIB
ATTN RDRL CIO MT TECHL PUB
ATTN RDRL D
ADELPHI MD 20783-1197

TOTAL: 18 (16 HCS, 1 CD, 1 ELECT)

INTENTIONALLY LEFT BLANK.